# HW_Data_Visualization

Kittipoom Bank

2023-07-02

## HW 1 - Analyze Data "Diamonds" [5 Charts]

**Load Library**

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

**Understanding Data**

```
str(diamonds)
```

```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

**Include with 10 Parameter & 53,930 Data**

**Data had 3 Types**

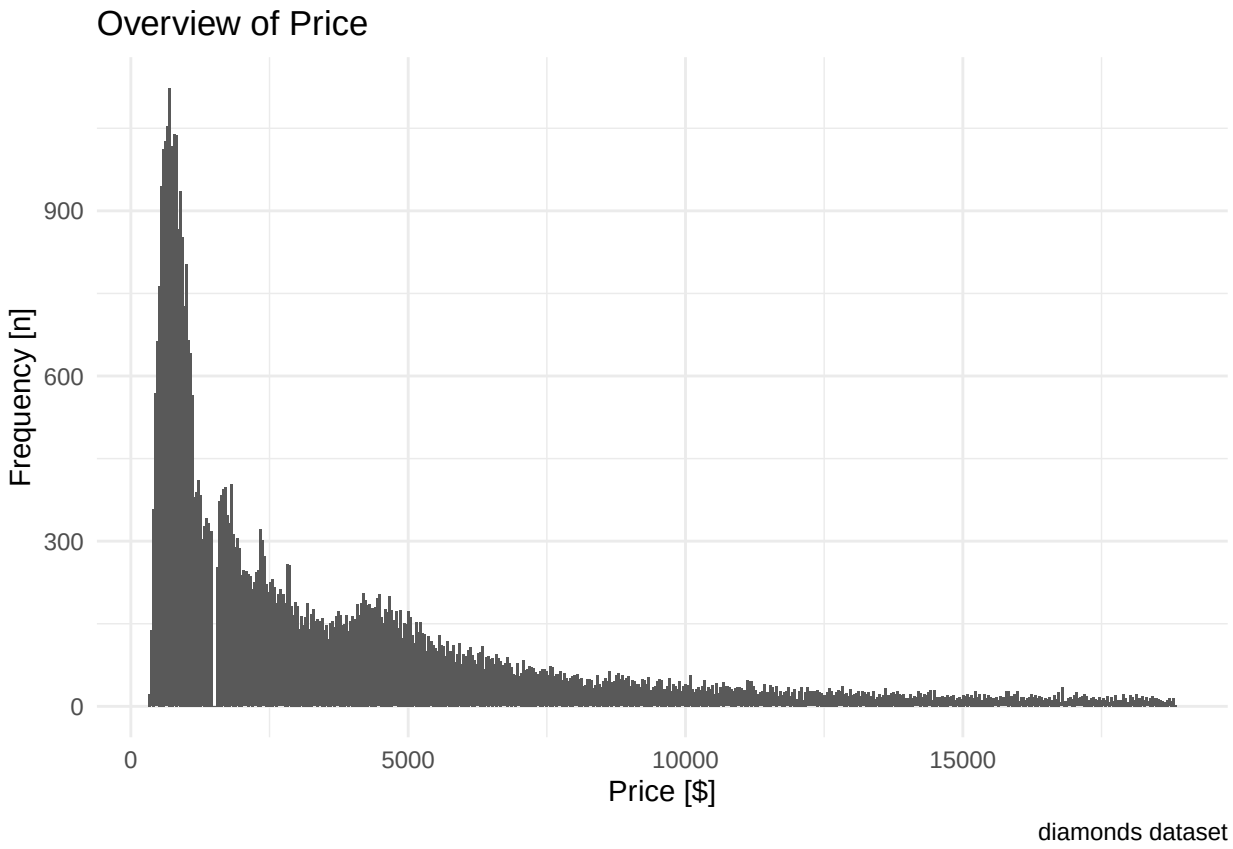
**(1) Ordered Factor : [cut / color / clarity]**

**(2) Numeric : [carat / depth / table / x / y / z]**

**(3) Integer : [price]**

**1: Overview of Price**

```
ggplot(diamonds,
            aes(price)) +
        geom_histogram(bins=500) +
        theme_minimal() +
        labs(
              title = "Overview of Price",
              x = "Price [$]",
              y = "Frequency [n]",
        caption = "diamonds dataset"
        )
```




```
diamonds %>%
  summarise(mean_price = mean(price),
            min_price = min(price),
            max_price = max(price),
            n = n())
```

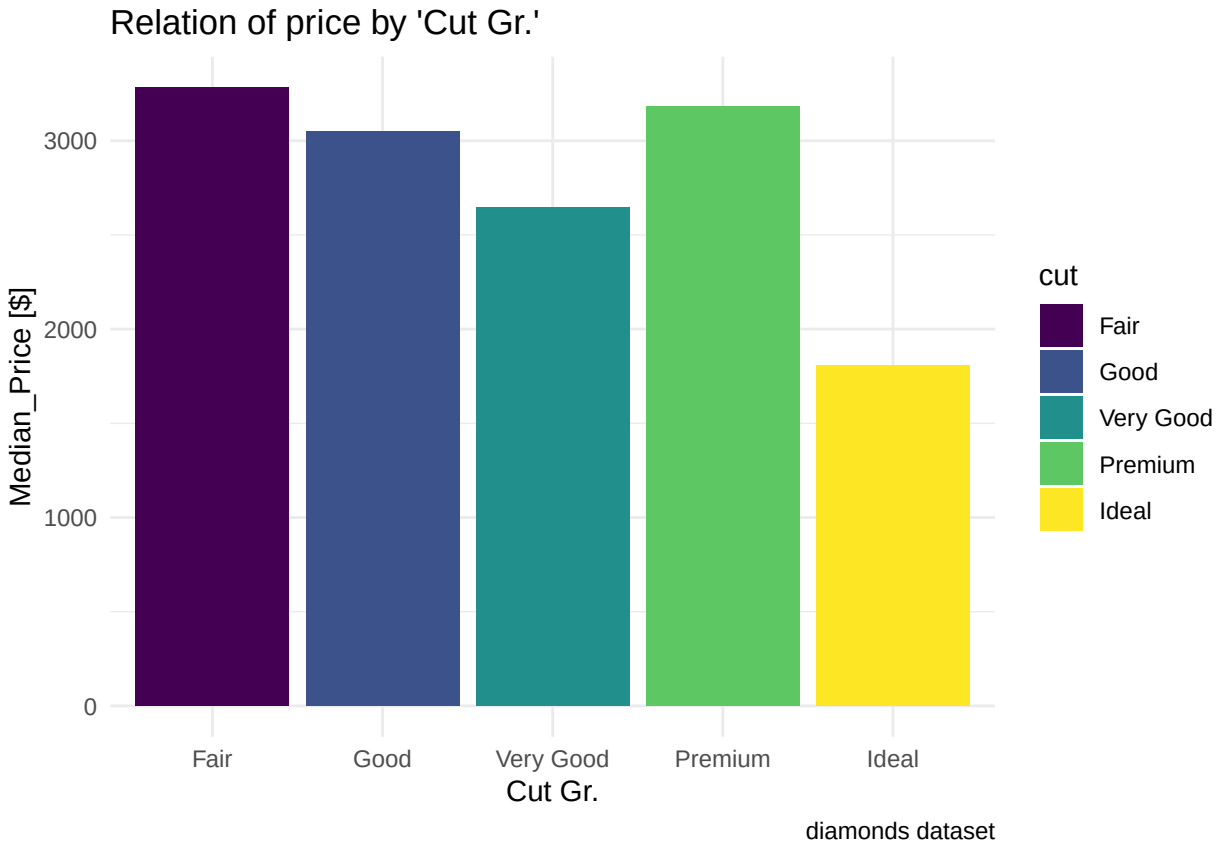
```
## # A tibble: 1 x 4
##   mean_price min_price max_price     n
##        <dbl>     <int>     <int> <int>
## 1      3933.       326     18823 53940
```

**Price is between : 326 - 18,823 $**

**By mean of price is 3,933 $**

**2: Relation of price by "Cut Gr."**

```
diamonds %>%
  group_by(cut) %>%
  summarise(
    med_price = median(price)
  ) %>%
  ggplot(aes(cut, med_price, fill=cut)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Relation of price by 'Cut Gr.'",
    x = "Cut Gr.",
    y = "Median_Price [$]",
  caption = "diamonds dataset"
  )
```




```
diamonds %>%
  group_by(cut) %>%
  summarise(mean_price = mean(price),
            min_price = min(price),
            max_price = max(price),
            n = n())
```

```
## # A tibble: 5 x 5
##   cut       mean_price min_price max_price     n
##   <ord>          <dbl>     <int>     <int> <int>
## 1 Fair           4359.       337     18574  1610
```

```
## 2 Good           3929.       327     18788  4906
## 3 Very Good      3982.       336     18818 12082
## 4 Premium        4584.       326     18823 13791
## 5 Ideal          3458.       326     18806 21551
```

**Classify by Median price in each group of cut, Rank from High Price (Top 3)**

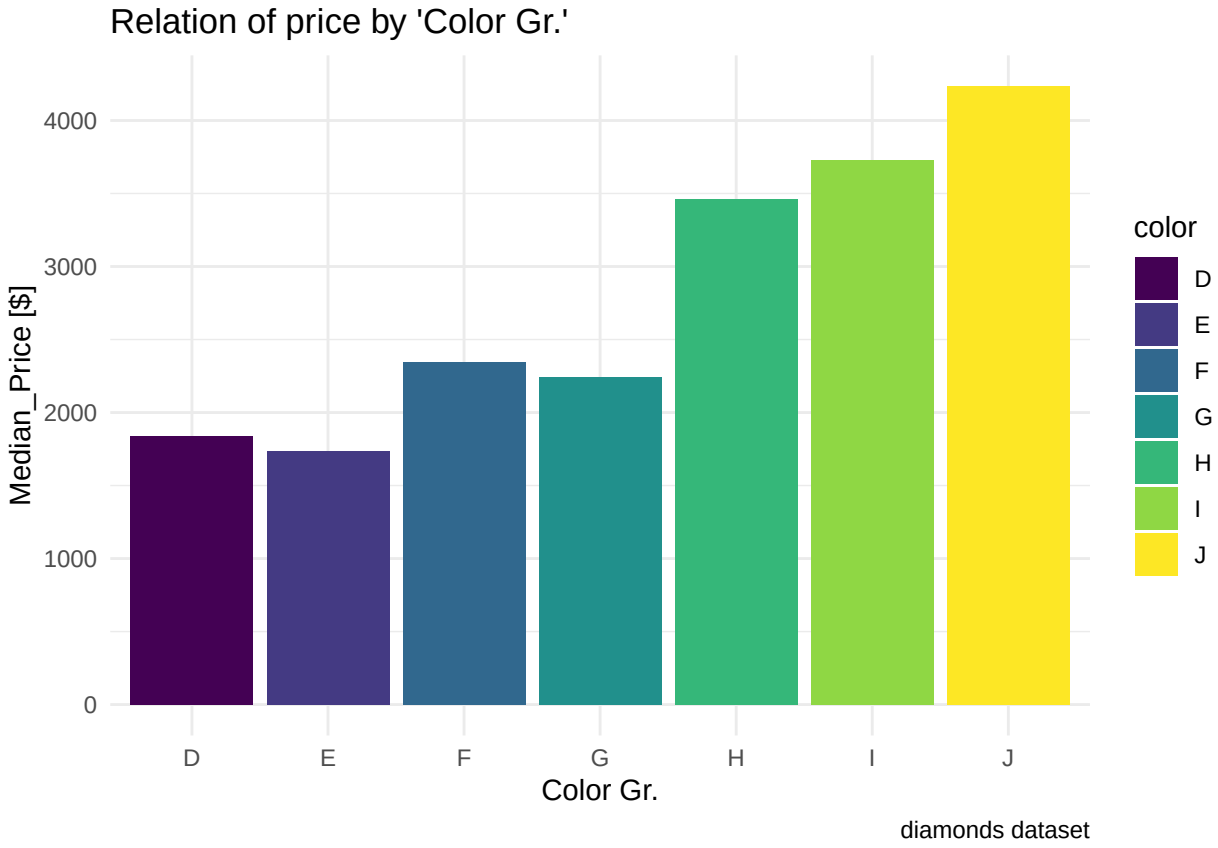
**(1) Premium [4,584 $]**

**(2) Fair [4,359 $]**

**(3) Very Good [3,982 $]**

**3: Relation of price by "Color Gr."**

```
diamonds %>%
  group_by(color) %>%
  summarise(
    med_price = median(price)
  ) %>%
  ggplot(aes(color, med_price, fill=color)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Relation of price by 'Color Gr.'",
    x = "Color Gr.",
    y = "Median_Price [$]",
  caption= "diamonds dataset"
  )
```




```
diamonds %>%
  group_by(color) %>%
  summarise(mean_price = mean(price),
            min_price = min(price),
            max_price = max(price),
            n = n())
```

```
## # A tibble: 7 x 5
##   color mean_price min_price max_price     n
##   <ord>      <dbl>     <int>     <int> <int>
## 1 D          3170.       357     18693  6775
## 2 E          3077.       326     18731  9797
## 3 F          3725.       342     18791  9542
## 4 G          3999.       354     18818 11292
## 5 H          4487.       337     18803  8304
## 6 I          5092.       334     18823  5422
## 7 J          5324.       335     18710  2808
```

**Classify by Median price in each group of color, Rank from High Price (Top 3)**

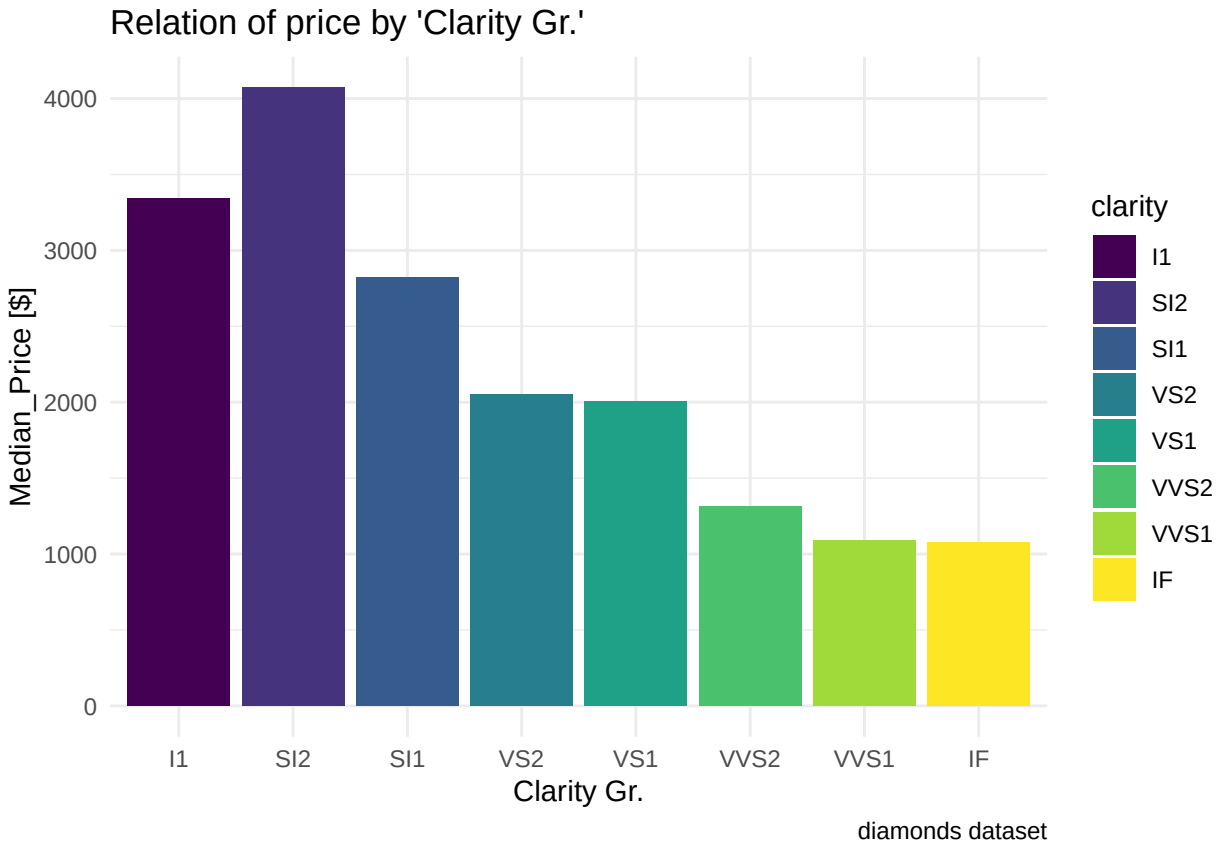
**(1) J [5,324 $]**

**(2) I [5,092 $]**

**(3) H [4,487 $]**

**4: Relation of price by "Clarity Gr."**

```
diamonds %>%
  group_by(clarity) %>%
  summarise(
    med_price = median(price)
  ) %>%
  ggplot(aes(clarity, med_price, fill=clarity)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Relation of price by 'Clarity Gr.'",
    x = "Clarity Gr.",
    y = "Median_Price [$]",
    caption = "diamonds dataset"
  )
```




```
diamonds %>%
  group_by(clarity) %>%
  summarise(mean_price = mean(price),
            min_price = min(price),
            max_price = max(price),
            n = n())
```

```
## # A tibble: 8 x 5
##   clarity mean_price min_price max_price     n
##   <ord>        <dbl>     <int>     <int> <int>
## 1 I1           3924.       345     18531   741
```

```
## 2 SI2          5063.       326     18804  9194
## 3 SI1          3996.       326     18818 13065
## 4 VS2          3925.       334     18823 12258
## 5 VS1          3839.       327     18795  8171
## 6 VVS2         3284.       336     18768  5066
## 7 VVS1         2523.       336     18777  3655
## 8 IF           2865.       369     18806  1790
```

**Classify by Median price in each group of clarity, Rank from High Price (Top 3)**

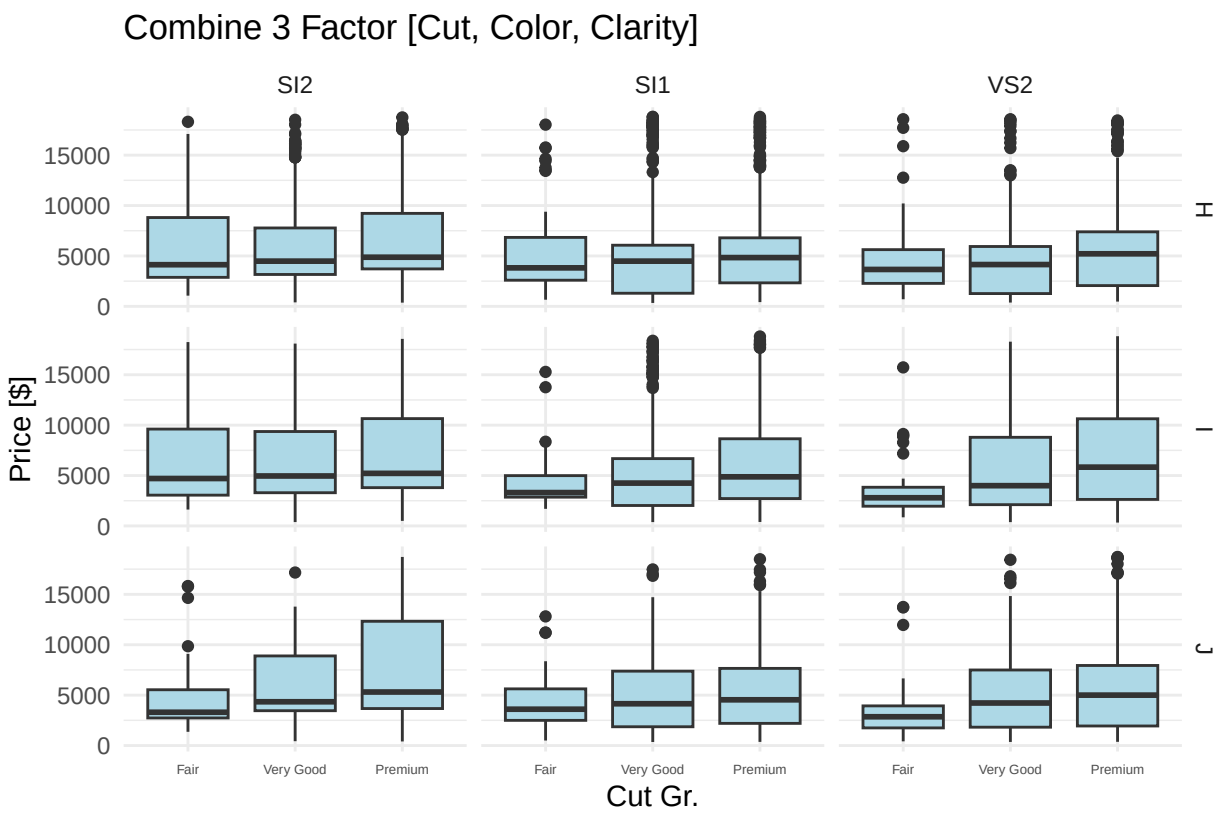
**(1) SI2 [5,063 $]**

**(2) SI1 [3,996 $]**

**(3) VS2 [3,925 $]**

**5: Combine 3 Factor (By Each Top 3) [Cut, Color, Clarity]**

```
diamonds %>%
  filter(diamonds$cut %in% c("Premium","Fair","Very Good") &
         diamonds$color %in% c("J","I","H") &
         diamonds$clarity %in% c("SI2","SI1","VS2")
         ) %>%
  ggplot(aes(cut, price)) +
  geom_boxplot(fill="light blue") +
  facet_grid(color ~ clarity) +
  theme_minimal() +
  labs(
    title = "Combine 3 Factor [Cut, Color, Clarity]",
    x = "Cut Gr.",
    y = "Price [$]",
    caption = "diamonds dataset"
  ) + theme(axis.text.x = element_text(size = 5))
```




```
diamonds %>%
  select (price, cut, color, clarity) %>%
  filter(diamonds$cut == "Premium" &
           diamonds$color == "I" &
           diamonds$clarity == "VS2") %>%
  summarise(mean_price = mean(price),
            min_price = min(price),
            max_price = max(price),
            n = n())
```

```
## # A tibble: 1 x 4
##   mean_price min_price max_price     n
##        <dbl>     <int>     <int> <int>
## 1      7156.       334     18823   315
```

**After Combine Top 3 in each Factor,**

**From Box Plot can observe High of Mean Price [7,156 $] in Group of**

**- Cut : Premium**

**- Color : I**

**- Clarity : VS2**